2019, March 4th-14th **John Von Neumann Institute**

Master QCF

Time series/Forecasting, course: 16 h + exercises: 8h.

Forecasting discipline is an issue of Statistics. Indeed, the aim is to answer the following kind of problem: a system X is evolving in time, it is observed and one would like to predict the future. Exemple: we can try to fill "holes" in a time series (missing data). Generally, underlies a modeling problem: it is to find the mathematical "model" that realizes the better connection between a variable and the time.
The methods are multiple. The principle is to find a mathematical modeling: for instance the series X is to forecast as a function of time. Given the available observations, we try the "best" function f (the optimality criterion depending on the method) such as $X \approx f(t)$ where t is time. Namely, we consider that the observations are a set $(X(t-i), i = 1, \cdots, n)$. This course presents three types of methods.

- The **Smoothing** (Brown, Holt and Winters, about the sixties) corresponds to the intuitive idea of "smoothing" the curve obtained using points observed for a smooth curve; smoothing provides $X(t)$ in terms of the past of X.

- **Linear regression,** really simple statistical method.

- Processes **ARMA, ARIMA, SARIMA** (Box and Jenkins): sophisticated methods, where is exhibited a linear function of $X(t)$ and its past values $X(t-i),\ i = 1, \cdots, n$.

Depending on the cases, one or the other of these methods are more suitable. We can not exclude one of them a priori. In the same study, it is convenient to use them each others and then compare their respective performances before fixing our choice. A selection criterion is obviously the quality of the forecast. Each method proposes statistical tests that allow to judge the quality of fit (between the curve obtained and the observations). An empirical way could be added: to reserve some "witnesses spots" and to do the study, excluding them, and judging the error on witnesses.

For the concrete use of these methods it is recommended to use the free software "R":

https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

The terminal test will be such a study on concrete data, with proposed models and their comparison.

1 Smoothing

(cf. Chapter IV, Gouriéroux-Monfort, [6])

1.1 Simple exponential smoothing

1.1.1 Definition, principle

This method leads to estimate the values of the series at times $T+k$ as function of the past values, exponentially, meaning:

$$\forall k \geq 1,\ \hat{X}_{T+k} = (1-\beta)\sum_{i=0}^{T-1} \beta^i x_{T-i} \tag{1}$$

when T observations x_i are available, $\beta \in]0,1[$. The interpretation is the following: x_{T-j} has less influence as j is high (more past).

In case of β close to 1, immediate past is less important than deep past; the forecasting is rigid; In case of β close to 0, immediate past is more important than deep past, the forecasting is flexible.

Warning: the forecasting is constant in the future: it is smoothing somehow "horizontal".

1.1.2 Update

Not to have to recalculate the total sum of formula (1), we have an " updating" formula:

$$\hat{X}_{T+1} = (1-\beta)x_T + \beta\hat{X}_T \tag{2}$$

The proof is simple: simply apply the formula (1) to $k=1$ and $k=0$ but with $T-1$ observations instead of T and to the linear combination $\hat{X}_{T+1} - \beta\hat{X}_T$ to find (2). We can interpret (2) as following:

- either let us do the center of gravity between the last forecast and the new observation,
- or in such similar form $\hat{X}_{T+1} = \hat{X}_T + (1-\beta)(x_T - \hat{X}_T)$ we add a weighting of innovation to the previous forecast.

1.1.3 Interpretation

Suppose we seek the *better constant* a to fit the series at a constant using the least squares, but with an exponential weighting, i.e., we try to minimize the function:

$$F : a \mapsto \sum_{i=0}^{T-1} \beta^i (x_{T-i} - a)^2.$$

The minimum of F is reached at

$$\tilde{a} = \frac{1-\beta}{1-\beta^T}\sum_{i=0}^{T-1}\beta^i x_{T-i},$$

i.e. roughly speaking, $\hat{X}_{T+1}$ is little different of $\tilde{a}$ when T is high enough (meaning a lot of observations). Thus, the interpretation of $\hat{X}_{T+1}$ is the better constant fitting the whole (β^i)-weighted series. This allows to conclude that this method is completely inappropriate in case of trending or seasonality. These points will be the subject of the following sections.

1.1.4 How to choose the constant β

Recall

- $\beta = 1-\varepsilon$, rigid, dependence from past,
- $\beta = \varepsilon$, flexible forecasting, low dependence on the past.

But there is a more objective criterion for choosing this smoothing constant: one chooses β which minimizes errors made replacing the observations by their prediction, namely, for $t = 1, \cdots, T-1,\ x_{t+1}$ replaced by $\hat{X}_{t+1} = (1-\beta)\sum_{i=0}^{t-1}\beta^i x_{t-i}$. So the aim is to minimize on the interval $[0,1]$ the application :

$$F : \beta \mapsto \sum_{t=1}^{t-1}(x_{t+1} - (1-\beta)\sum_{i=0}^{t-1}\beta^i x_{t-i})^2.$$

In the general case, the computations are tedious and inextricable. Nevertheless, there is necessarily a solution since F is a continuous function on a compact. The study of its variations is complicated. Therefore a special case could be solved when x_k are the values of the random variables X_k, stationary centered 2-order series with the covariance function $\gamma(k) = \rho^{|k|}, |\rho| < 1$. And rather than seek to minimize the sum of observed squared deviations, we could minimize their mean:

$$G : \beta \mapsto E[(X_{t+1} - (1-\beta)\sum_{i=0}^{t-1}\beta^i X_{t-i})^2].$$

It yields

$$G(\beta) = 1 - 2(1-\beta)\sum_{i=0}^{t-1}\beta^i\rho^{i+1} + (1-\beta)^2\sum_{i,j\leq t-1}\beta^{i+j}\rho^{|i-j|}.$$

We suppose t high enough and we compute the last term:

$$\sum_{i,j}\beta^{i+j}\rho^{|i-j|} = \sum_{i=0}^{\infty}\beta^{2i} + 2\sum_{i=0}^{\infty}\sum_{j=i+1}^{\infty}\beta^{i+j}\rho^{|i-j|} = \frac{1}{1-\beta^2} + \frac{2\rho\beta}{(1-\beta^2)(1-\rho\beta)}$$

$$\text{so } G(\beta) = 1 - 2\frac{(1-\beta)\rho}{(1-\rho\beta)} + \frac{1-\beta}{1+\beta} + \frac{2\rho\beta(1-\beta)}{(1+\beta)(1-\rho\beta)} = 2\frac{1-\rho}{(1+\beta)(1-\rho\beta)}.$$

The optimum is depending on ρ and on the position of $\frac{1-\rho}{2\rho}$ with respect to 0 and 1. Let be the logarithmic derivative of G with respect to β : the denominator is positive and the numerator is $2\rho\beta - 1 + \rho$.

(i) if $0 < \frac{1-\rho}{2\rho} < 1$, i.e. $\rho > 1/3$, the optimum is $\hat{\beta} = \frac{1-\rho}{2\rho}$,

(ii) if $\frac{1-\rho}{2\rho} \geq 1$, i.e. $\rho \leq 1/3$, the optimum is $\hat{\beta} = 1$.

This means that in the case of a low correlation, this is not a good method: the " best" forecast is $\hat{X} = 0$ which on the one hand is not very interesting and on the other hand certainly gives rise to very large errors. We must seek in this case another method.

In case (i) (good correlation), the minimum value of G when $\hat{\beta} = \frac{1-\rho}{2\rho}$ is:

$$G(\frac{1-\rho}{2\rho}) = \frac{8\rho(1-\rho)}{(1+\rho)^2}.$$

In practice, if we trace the family of curves representing $G_\rho(\beta)$ we find that " good " β values are in the range of 0.7 to 0.8 corresponding to values of ρ around 0.4.

Another good and solvable example is the case when the auto-covariance function γ is zero for $|k|$ large enough, for example, the series in Exercise 4 of the sheet 1. In such a case we can find an optimal β.

1.2 Double exponential smoothing=Lissage exponentiel double

1.2.1 Definition, principle

This method is convenient when a linear trend is possible. The principle is to fit the series to a line: $a_1 + (t-T)a_2$ instead of a constant:

$$\hat{X}_{T+k} = \hat{a}_1(T) + k\hat{a}_2(T)$$

where T is the length of the forecasting. We look for the constants $\hat{a}_1(T)$ and $\hat{a}_2(T)$ which minimize the following application:

$$F : (a_1, a_2) \mapsto \sum_{j=0}^{T-1} \beta^j (x_{T-j} - a_1 + a_2 j)^2$$

meaning the quadratic mean (weighted exponentially by β^j) of the errors which are resulting of the replacement of observation x_{T-j} by the estimate with a trend: $a_1 - a_2 j$.

Using standard formulas $\sum_{j\geq 1} j\beta^j = \frac{\beta}{(1-\beta)^2}$ or $\sum_{j\geq 1} j^2\beta^j = \frac{\beta(1+\beta)}{(1-\beta)^3}$ we deduce both partial derivatives of the convex function F, meaning:

$$\begin{aligned}
-\frac{1}{2}\nabla_1 F &\simeq \sum_{j=0}^{T-1} \beta^j x_{T-j} - a_1 \frac{1}{1-\beta} + a_2 \frac{\beta}{(1-\beta)^2}, \\
-\frac{1}{2}\nabla_2 F &\simeq \sum_{j=0}^{T-1} \beta^j j x_{T-j} - a_1 \frac{\beta}{(1-\beta)^2} + a_2 \frac{\beta(1+\beta)}{(1-\beta)^3}
\end{aligned}$$

asuming once again T close to infinity to simplify the computations. Let be:

$$\begin{aligned} S_1(T) &= (1-\beta)\sum_{j=0}^{T-1}\beta^j x_{T-j} \text{ named as "smoothed series" ;} \\ S_2'(T) &= (1-\beta)\sum_{j=0}^{T-1}\beta^j j x_{T-j}, \end{aligned}$$

after tedious but straightforward computations, the unique pair canceling the gradient F is:

$$\begin{aligned} \hat{a}_1(T) &= (1+\beta)S_1(T) - (1-\beta)S_2'(T), \\ \hat{a}_2(T) &= (1-\beta)S_1(T) - \frac{(1-\beta)^2}{\beta}S_2'(T). \end{aligned}$$

1.2.2 Update

For updating the coefficients $a_i(T)$ remark that:

$$S_1(T) - \beta S_1(T-1) = (1-\beta)x_T,$$

and we deduce the simple updating:

$$S_1(T) = (1-\beta)x_T + \beta S_1(T-1).$$

The updating of the sum S_2 is more difficult; we introduce the series " doubly" smoothed:

$$S_2(T) = (1-\beta)\sum_{j=0}^{T-1}\beta^j S_1(T-j)$$

Exercise: prove the relation between S_2 and S_2':

$$S_2'(T) = \frac{1}{1-\beta}S_2(T) - S_1(T).$$

Updating this new sum is a little bit simpler. We prove:

$$S_2(T) = \beta S_2(T-1) + (1-\beta)^2 x_T + \beta(1-\beta)S_1(T-1).$$

Then we deduce (once again after tedious but straightforward computations!!) the updatings:

$$\begin{aligned} \hat{a}_1(T) &= x_T(1-\beta^2) + \beta^2[\hat{a}_1(T-1) + \hat{a}_2(T-1)], \\ \hat{a}_2(T) &= x_T(1-\beta)^2 + \hat{a}_2(T-1) - (1-\beta)^2[\hat{a}_1(T-1) + \hat{a}_2(T-1)]. \end{aligned}$$

1.2.3 Procedure

look for routines in software R

Statgraphics sofware provides some routines: **BROWN**, three options : simple (no trend), linear(meaning “double" or linear trend) and quadratic (which could be named “triple" and corresponding to a quadratic trend. We need the smoothing constant, here named: “smoothing constant alpha"...

1.3 Generalized Exponential Smoothing

We will try to fit the observations to more sophisticated functions, more than the constant function or the line, in particular to take in account the seasonality (periodic functions). The first to do that was Brown (1962) who proposes the following tool.

1.3.1 State-transition matrices

Definition 1.1. *We say that* $f : Z \mapsto \mathbb{R}^n$ *is with* **State-transition matrix** *if there exists a matrix* A *with non null determinant and such that:*

$$f(t) = Af(t-1),\ \forall t \in Z.$$

The principle of the generalized exponential smoothing is to fit the series X_t with $\varphi(t-T)$ where $\varphi(t) = \sum_{i=1}^{n} a_i f_i(t)$. Look at some examples:

(a) $\varphi(t) = a$ is Subsection 1.1, simple smoothing. It is obtained with the constant function $f(t) = 1$ and the matrix $A = 1$ in $1-$dimension. Then$\hat{X}_{T+k} = \varphi(k) = \hat{a}(T)$.

(b) $\varphi(t) = a_1 + a_2 t$ is Subsection 1.2, double smoothing. It is obtained with the function $f(t) = (1, t)$ and the matrix A in $2-$dimension:

$$\begin{matrix} 1 & 0 \\ 1 & 1 \end{matrix}$$

Actually the determinant of this matrix is non null. Thus, $\hat{X}_{T+k} = \varphi(k) = \hat{a}_1(T) + \hat{a}_2(T)k$.

(c) $\varphi(t) = a_1 \sin \omega t + a_2 \cos \omega t$, is obtained with the function $f(t) = (\sin \omega t, \cos \omega t)$ and the matrix A in $2-$dimension:

$$\begin{matrix} \cos\omega & \sin\omega \\ -\sin\omega & \cos\omega \end{matrix}$$

Actually the determinant of this matrix is non null. Thus, $\hat{X}_{T+k} = \varphi(k) = \hat{a}_1(T) \sin \omega k + \hat{a}_2(T) \cos \omega k$.

(d) $\varphi(t) = ae^{\alpha t}$, is obtained with the function$f(t) = e^{\alpha t}$ and the matrix $A = e^{\alpha}$ in $1-$dimension. Actually the determinant of this matrix is non null. Thus, $\hat{X}_{T+k} = \varphi(k) = \hat{a}(T)e^{\alpha k}$.

1.3.2 The method

We forecast X_t with the scalar product in $\mathbb{R}^n$, $\varphi(t-T) = \langle a, f(t-T)\rangle$. The function f being fixed (it is the "form" of the fitting, the smoothing) we look for an optimization with respect to a, meaning to minimize the application, computed on the available observations:

$$G : a \mapsto \sum_{j=0}^{T-1} \beta^j (x_{T-j} - \langle a, f(-j)\rangle)^2,$$

so we have to cancel the gradient of the function G, a convex function:

$$-\frac{1}{2}\nabla_i G = \sum_{j=0}^{T-1} f^i(-j)\beta^j (x_{T-j} - \langle a, f(-j)\rangle) = 0.$$

Let be Y the vector $(x_T, \cdots, x_T - j, \cdots, x_1)$, F_β the matrix with general coefficient $f^i(-j)\beta^j$ arrow i and column j. The above system admits the matrix writing:

$$F_\beta.F^t.a = F_\beta.Y.$$

Let be the optimal vector:

$$\hat{a}(T) = (F_\beta.F^t)^{-1} F_\beta.Y.$$

To be simpler, as above , we suppose T high enough in $F_\beta F^t$ i.e. $T \sim \infty$ and since $0 < \beta < 1$ all the series are convergent, so the matrix $F_\beta F^t$ with general term $\sum_{k\geq 0} f^i(-k)\beta^k f^j(-k)$ does not depend on T.
Exercise: to solve the examples (c) and (d).

1.3.3 Update

We recall $\hat{a}(T) = (F_\beta.F^t)^{-1} \sum_{j=0}^{T-1} \beta^j f(-j) x_{T-j}$.
In the expression $\sum_{j=0}^{T-1} \beta^j f(-j) x_{T-j}$ we can focus on the last observation:

$$\sum_{j=0}^{T-1} \beta^j f(-j) x_{T-j} = f(0)x_T + \sum_{j=1}^{T-1} \beta^j f(-j) x_{T-j} = f(0)x_T + \beta \sum_{j=0}^{T-2} \beta^j f(-j-1) x_{T-1-j}.$$

But the hypothesis implies $f(t) = Af(t-1)$, so $f(-j) = Af(-j-1)$ and:

$$\hat{a}(T) = (F_\beta.F^t)^{-1} f(0) x_T + \beta (F_\beta.F^t)^{-1} A^{-1} (F_\beta.F^t) \hat{a}(T-1).$$

This means that $\hat{a}(T)$ could be written as $g x_T + G\hat{a}(T-1)$ with $g = (F_\beta.F^t)^{-1} f(0)$ and $G = \beta (F_\beta.F^t)^{-1} A^{-1} (F_\beta.F^t)$, and these matrices do not depend on time, so they are computable from the beginning.

As previously we have some updating formulas which stress the so called "innovation":

$$\hat{a}(T) = (g.f^t(1) + G)\hat{a}_{T-1} + g(x_T - \hat{X}_T(T-1)).$$

Exercise 4, sheet 2.
The most important problem, given the observations graph, is the recognition of the smoothing curve deduced from the function f. It is less reliable than Box and Jenkins' methods that we will see in the third chapter, but, nevertheless, it may be useful.

1.4 Holt et Winters' methods

They are based on Winters (1960, cf. [12]) seminal work, or Harrison [8].

1.4.1 No seasonality

The principle is the fitting of X_t on $a_1 + (t-T)a_2$ but with different updating

$$\begin{aligned} &\hat{a}_1(T) = (1-\alpha)x_T + \alpha[\hat{a}_1(T-1) + \hat{a}_2(T-1)],\ 0<\alpha<1 \\ &\hat{a}_2(T) = (1-\gamma)[\hat{a}_1(T) - \hat{a}_1(T-1)] + \gamma\hat{a}_2(T-1),\ 0<\alpha,\gamma<1. \end{aligned} \tag{3}$$

The initial constants are arbitrary, but the practice advises to take:

$$\hat{a}_1(2) = x_2\ ;\ \hat{a}_2(2) = x_2 - x_1.$$

This can be understood as follows: $\hat{a}_1(T)$ is the "forecast" $\hat{X}_T$ with T observations, but the observation is x_T. Otherwise, with $T-1$ observations, $\hat{X}_T = \hat{a}_1(T-1) + \hat{a}_2(T-1)$. The update $\hat{a}_1$ is the center of gravity between these two possibilities of $\hat{X}_T$.
Similarly for $\hat{a}_2$, we can "forecast" x_{T+1} either with T observations, and it is $\hat{a}_1(T)+\hat{a}_2(T)$, or with $T-1$, and it is $\hat{a}_1(T-1)+2\hat{a}_2(T-1)$. In case of both identical forecasts, it implies an estimate of $\hat{a}_2$ equal to $-\hat{a}_1(T) + \hat{a}_1(T-1) + 2\hat{a}_2(T-1)$.
Concerning x_{T-1}, it is "forecasted" either with T observations, so by $\hat{a}_1(T) - \hat{a}_2(T)$, or with $T-1$ observations, so by $\hat{a}_1(T-1)$.
In case of both identical forecasts, it implies an estimate of $\hat{a}_2$ equal to $\hat{a}_1(T) - \hat{a}_1(T-1)$. Choosing the β-barycenter between these two cases, we get the proposition with $\gamma = 2\beta$. A similar raisoning with two observations warrants the proposed initialization.

Notice that using $\hat{X}_T = \hat{a}_1(T-1) + \hat{a}_2(T-1)$, (3) could be written as:

$$\begin{aligned} &\hat{a}_1(T) = (1-\alpha)(x_T - \hat{X}_T) + \hat{a}_1(T-1) + \hat{a}_2(T-1), \\ &\hat{a}_2(T) = (1-\gamma)(1-\alpha)[x_T - \hat{X}_T] + \hat{a}_2(T-1) \end{aligned} \tag{4}$$

to highlight the dependence on the last observation.

In this method, there is **two** constants which allows greater flexibility of use.

Exercise: Compare these update formulas from those obtained in the paragraph 1.2.2: Exercise 4 Sheet 2.

The interpretation is similar to the previous case: if these constants are close to 1, forecasts are " smooth" and depend heavily on the past. Thus, the forecast is:

$$\hat{X}_T(k) = \hat{a}_1(T) + k\hat{a}_2(T).$$

We could choose the constants α and γ minimizing the following function, calculated on available observations:

$$(\alpha,\gamma) \mapsto \sum_{t=1}^{T-1}(x_{t+1} - \hat{X}_{t+1})^2 = \sum_{t=1}^{T-1}(x_{t+1} - \hat{a}_1(T)(\alpha,\gamma) - (t+1-T)\hat{a}_2(T)(\alpha,\gamma))^2.$$

1.4.2 Additive seasonality

We look for a fitting of the series with the function:

$$t \mapsto a_1 + (t - T)a_2 + S_t$$

where there is a trend a_2 but also a seasonal (here additive) factor S_t. The authors propose update formulas following, where s is the number of "season "', e.g. 12 monthly data, 4 for quarterly data, etc.

$$\hat{a}_1(T) = (1-\alpha)(x_T - \hat{S}_{T-s}) + \alpha[\hat{a}_1(T-1) + \hat{a}_2(T-1)], \quad (5)$$

$$\hat{a}_2(T) = (1-\gamma)[\hat{a}_1(T) - \hat{a}_1(T-1)] + \gamma\hat{a}_2(T-1), \quad (6)$$

$$\hat{S}_T = (1-\delta)[x_T - \hat{a}_1(T)] + \delta\hat{S}_{T-s}, \quad (7)$$

where the constants $\alpha, \gamma, \delta \in]0,1[$. These formulas are similar to (3) where x_T is replaced by its 'seasonalized' value; the second is the same; the third is natural enough: we weight between the previous value and $x_T - \hat{a}_1(T)$ matching prediction relationship given below:

$$\hat{X}_T(0) = \hat{a}_1(T) + \hat{S}_{T-s}.$$

Finally the forecast is:

$$\hat{X}_{T+k}(T) = \hat{a}_1(T) + k\hat{a}_2(T) + \hat{S}_{T+k-is},\ (i-1)s < k \leq is,\ \forall i.$$

The practical problem is still the choice of the smoothing constants, here α, γ, δ.

Moreover, it is necessary to initialize these constants. Gouriéroux and Monfort propose the following set of initial constants based on the need to have a priori estimates of $\hat{S}_i, i = 1, \ldots, s$ since, by building, the recurrence begins only at $T = s + 1$ and requires the data of $\hat{S}_i, i = 1, \ldots, s$:

$$\begin{aligned}
\hat{a}_1(3) &= 1/8x_1 + 1/4x_2 + 1/4x_3 + 1/4x_4 + 1/8x_5 \\
\hat{a}_1(4) &= 1/8x_2 + 1/4x_3 + 1/4x_4 + 1/4x_5 + 1/8x_6 \\
\hat{a}_2(4) &= \hat{a}_1(4) - \hat{a}_1(3)\ ;\ \hat{a}_1(2) = \hat{a}_1(3) - \hat{a}_2(4) \\
\hat{a}_1(1) &= \hat{a}_1(3) - 2\hat{a}_2(4)\ ;\ \hat{S}_i = x_i - \hat{a}_1(i).
\end{aligned}$$

1.4.3 Multiplicative seasonality

Here we look for a fitting of the series with the function:

$$t \mapsto (a_1 + (t - T)a_2)S_t$$

where a_2 is the trend coefficient but the seasonal factor S_t is multiplicative. The authors propose the following update formulas, where s is the "season " number, e.g. 12 for monthly data, 4 for quarterly data, etc.

$$\hat{a}_1(T) = (1-\alpha)\frac{x_T}{\hat{S}_{T-s}} + \alpha[\hat{a}_1(T-1) + \hat{a}_2(T-1)], \quad (8)$$

$$\hat{a}_2(T) = (1-\gamma)[\hat{a}_1(T) - \hat{a}_1(T-1)] + \gamma\hat{a}_2(T-1), \quad (9)$$

$$\hat{S}_T = (1-\delta)\frac{x_T}{\hat{a}_1(T)} + \delta\hat{S}_{T-s}, \quad (10)$$

where the constants $\alpha, \gamma, \delta \in]0,1[$. Finally the forecast is:

$$\hat{X}_T(k) = [\hat{a}_1(T) + k\hat{a}_2(T)]\hat{S}_{T+k-is},\ (i-1)s \leq k \leq is,\ \forall i.$$

Once again, the practical problem is the choice of the smoothing constants, here α, γ, δ.

1.4.4 Procedure

R Software provides routines **HoltWinter**. Look at "A little Book of R for Times Series".

2 Regression

2.1 Introduction

The principle of regression applied to forecasting is as follows: let numerical data be indexed by time, meaning a set of points; the goal is to "fit" these points by a curve and thus to extrapolate the future (and possibly the past!), or to reconstruct missing data by interpolation of this curve. So, we have a series $(t_i, X_i)i = 1, \cdots, n$, in $\mathbb{R}^+ \times \mathbb{R}$ and we are looking for a function $f : \mathbb{R}^+ \to \mathbb{R}$ which "minimizes" the quantity $((X_i - f(t_i))i = 1, \cdots, n)$. Here the criterium is the "least squares criterium", meaning to minimize the application:

$$f \mapsto \|X - f(t)\|^2 = \sum_{i=1,\cdots,n} (X_i - f(t_i))^2.$$

The most common types of fitting are:

$$f(t) = a + bt; a + bt + ct^2; \text{ a } n \text{ degree polynomial,}$$

or for instance:

$$\begin{aligned} f(t) \quad = \quad & a\cos(\omega t + \phi), \text{ periodic function;} \\ & a\log t + b; \\ & \frac{1}{a + bt}, \text{ reciprocal function;} \\ & a.b^t; a.b^t + c, \text{ exponential and modified exponential function;} \\ & a.t^b; a.t^b + c, \text{ power and modified power functions;} \\ & \frac{a}{1 + b.c^t}, \text{ logistic function ;} \\ & \exp[a.b^t + c], \text{ Gompertz function.} \end{aligned}$$

Exercise : show which of these fittings can be reduced to linear regression by one or more appropriate variable changes.

The principle is to "guess at sight" the type of function to choose, according to the profile of the observed points; then to estimate the parameters by minimizing the quadratic difference; finally to validate the model by statistical tests on residuals, i.e. the random variables $(\varepsilon_i = X_i - f(t_i),\ i = 1, \cdots, n)$ on which we make the assumption that the law is a centered Gaussian law. So we do a Fisher test to know if ε are small enough.

2.2 Linear regression

It is the most used in practice, even if it is not necessarily the most efficient!!

Definition 2.1. *The* **regression line** *of Y with respect to X is the line $x \to a+bx$ where the parameters (a, b) minimize the quantity:*

$$F(a, b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2,$$

meaning the so called "least squares method".
The interpretation of this line is as follows: if we draw on a graph (x, y) the population points i with coordinates $\{(x_i, y_i)i = 1, ..., n\}$, this line is the one that passes as close as possible to all these points. Indeed, for every point i, the quantity $(y_i - a - bx_i)^2$ is the square of the vertical distance between this point and the line $x \to a + bx$.

2.2.1 Regression parameters

Above we defined function F, convex, differentiabledepending on two variables. Thus a point which cancels both partial derivatives is a minimum for F. The constants a and b are the solutions of the linear system:

$$\begin{aligned} \partial_a F &= -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0, \\ \partial_b F &= -2\sum_{i=1}^{n}(y_i - a - bx_i)x_i = 0. \end{aligned} \tag{11}$$

So

$$\begin{aligned} \sum_{i=1}^{n} y_i &= na + b\sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y_i &= a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2. \end{aligned}$$

After some computations, parameters a and b are:

$$\hat{b} = \frac{S_{x,y}}{\sigma_x^2} \ ; \ \hat{a} = \bar{y} - \bar{x}\hat{b},$$

where

$$S_{x,y} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y},$$

$\bar{x}$ and $\bar{y}$ denoting the empirical means of the variables X and Y; σ_x, σ_y are they empirical standard deviations. Moreover here we use the "covariance":

Definition 2.2. *The* **empirical covariance** *of the variables X and Y is:*

$$S_{xy} = \frac{1}{n}\sum_{i=1,...,n}(x_i - \bar{x}) \times (y_i - \bar{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y}.$$

The coefficients $\hat{a}$ are $\hat{b}$ are "estimated regression coefficients"; the line $y = \hat{a} + \hat{b}x$ is a trend line, fitting Y with respect to X.

Proposition 2.3. *We assume $\varepsilon_i = X_i - a - bt_i$ is a centered Gaussian random variable with variance σ^2; then the estimates $\hat{b}$ and $\hat{a}$ are too Gaussian random variables, respectively $\mathcal{N}(b, \frac{\sigma^2}{nVar(t)})$ and $\mathcal{N}(a, \frac{\sigma^2}{n}(1 + \frac{\bar{t}^2}{Var(t)}))$.*

Standard routines provide these estimates and their law.

2.2.2 Forecasting

We now can use this fitting line to forecast Y values using observed X values. Or to forecast X values using new times t_i:

$$y_{n+1} = \hat{a} + \hat{b}x_{n+1}, x_{n+1} = \hat{a} + \hat{b}t_{n+1}.$$

2.2.3 Correlation

The covariance value belongs to the interval $[-\sigma_x \times \sigma_y, \sigma_x \times \sigma_y]$. The following does not depend on the unit:

Definition 2.4. *The* **empirical correlation coefficient** *of the statistical variables X and Y is $\hat{\rho}_{x,y} := \frac{S_{xy}}{\sigma_x \times \sigma_y}$, or in temporal fitting case: $\hat{\rho}_{t,x} = \frac{S_{xt}}{\sigma_x \times \sigma_t}$.*

Remark that $\hat{\rho} \in [-1, -1]$.
Exercise: prove $\hat{\rho} \in [-1, -1]$ and moreover

$$\hat{\rho} = +1 \text{ ou } -1 \Leftrightarrow \forall i = 1, ..., n \ , \ y_i = a + bx_i,$$

meaning a perfect linear fitting.

Interpretation: more $\hat{\rho}^2$ is close to 1, better is the link between X and Y, the approximation of Y by $a + bX$, of X by $a + bt$, is "valuable".

2.2.4 Study of the residuals

Having the estimates of a et b, it remains the differences, the fitting errors.

Definition 2.5. *The* **residuals** *are the differences*

$$\varepsilon_i = x_i - \hat{a} - \hat{b}t_i.$$

These ones are supposed to be small since these are mistakes made when admitting the model $X = a + bt$. These residuals satisfy some properties:

- they are centered:

$$\sum_i \varepsilon_i = n\bar{x} - n\hat{a} - \hat{b}n\bar{t} = 0$$

using $\hat{a}$.

- they are non correlated with the t_i :

$$\sum_i \varepsilon_i(t_i - \bar{t}) = \sum_i (x_i - \hat{a} - \hat{b}t_i)t_i = -\frac{1}{2}\partial_b F = 0$$

using Equation (11).

- their variance is

$$s^2(\varepsilon) = \frac{1}{n}\sum_1^n \varepsilon_i^2 = \sigma_y^2(1-\hat{\rho}^2).$$

If this quantity is "too high", we can not accept the model. Indeed, we assume that the ε_i are the values taken by a Gaussian random variable.Thus, it makes possible locating aberrant values: the probability that the residuals are outside of the interval $[-2s(e), +2s(e)]$ is small ($\mathbb{P}\{|\varepsilon| > 2\sigma\} = 0.0456$), and it could be good to take a closer look at the corresponding points.

Example : Let Y be the son's size and X the father's size. The estimates are

$$a = 84.843\ ;\ b = 0.532\ ;\ \sigma_y^2 = 39.73\ ;\ \rho = 0.533\ ;\ s^2(e) = 28.44.$$

For $x = 165$cm, the average size of the son is estimated by $y = 0.532 \times 165 + 84.843 = 172.66$.

More specifically we get:

Proposition 2.6. *We asume that $\varepsilon_i = X_i - a - bt_i$ are centered Gaussian random variables with variance σ^2, the random variable $\frac{\sum_{i=1}^n \varepsilon_i^2}{\sigma^2}$ law is the χ^2_{n-2} law, and it is independent of the random variables $\hat{a}$ and $\hat{b}$.*

This result allows to get an unbiaised estimate of σ^2, meaning $\frac{\sum_{i=1}^n \varepsilon_i^2}{n-2}$, and a confident interval for this parameter; this is a way to measure the error. In the general case, we do not really know this parameter σ^2. On the other hand, we have the following result to control the estimates of a and b.

Proposition 2.7. *The law of the random variables*

$$T = \frac{\sigma_t(\hat{b}-b)}{\sqrt{\frac{s^2(\varepsilon)}{n-2}}} = \frac{\sqrt{n-2}Var(t)(\hat{b}-b)}{\sqrt{Var(t)Var(X) - cov^2(X,t)}},$$

and

$$S = \frac{\sigma_t(\hat{a}-a)}{\sqrt{\frac{s^2(\varepsilon)(Var(t)+\bar{t}^2)}{n-2}}} = \frac{\sqrt{n(n-2)}Var(t)(\hat{a}-a)}{\sqrt{\sum_i t_i^2}\sqrt{Var(t)Var(X) - cov^2(X,t)}},$$

is a Student$_{n-2}$ law.

2.2.5 Correlation coefficient test

There are statistical methods to know if the estimate of the correlation coefficient ρ is "significantly" small. Indeed, $\hat{\rho}$ is the value taken by an estimator, a random variable with a known law. Actually we show that the associated variable $F = (n-2)\frac{\hat{\rho}^2}{1-\hat{\rho}^2}$ law is a Fisher-Snédécor law of degrees of freedom $(1, n-2)$, which makes it possible to test the hypothesis $\rho = 0$ against $\rho \neq 0$. Thus we compute the value taken by F and we examine in the Fisher-Snedecor table if this value is small enough or not.

Exercise : $n = 63$; $\hat{\rho} = 0.533$; $F = 24.206$.
The probability that the random variable Fisher-Snedecor$_{1,61}$ is so huge is almost negligible: we can not accept that ρ is zero.
Be careful! : when n is large, the value of F is too large, and almost always significant! we then tend to reject the hypothesis $\rho = 0$..., maybe wrongly.

2.2.6 Regression of X with respect to Y

We can, in the same way that we try to adjust Y according to X, try to adjust X according to Y. By symmetry, we simply find another regression line:

$$x = a' + b'y, \text{ where } b' = \frac{S_{x,y}}{\sigma_y^2} \ , \ a' = \bar{x} - \bar{y}b',$$

and we notice that the two slopes b and b' are linked by the relation:

$$bb' = \rho_{x,y}^2$$

which means that the two lines are even closer than the coefficient correlation is closer to ± 1. Remind the test on the correlation coefficient, cf. (2.2.3).

2.2.7 The practice

Look at these routines in sofware R BUT, there is no regression in 'a Little Book of R for Time Series'. I will provide another booklet.
Using software Statgraphics, the command **REG** is the one for *Simple regression* which concerns this model. Obviously, we have an interest in reading carefully the manual ... The following screen shows where to name the variable to explain, then the explanatory variable (eg time). We indicate *linear* for *model* then the probabilities of confidence for the tests to do for validating the model obtained. After turning the procedure, we get a second screen with the digital outputs.

- *intercept* is the ordinate at x or $t = 0$,

- *slope* of the regression line.

After that, the standard deviation for $\hat{a}$ and $\hat{b}$ are given. These ones allow to produce confident ingterval, the value of Student variable, so we can test the hypotheses T or $S = 0$. (cf. Proposition 2.7).

- variance analysis: the value *residual* is the sum of squared errors, with degrees of freedom *Df* (here 152), then the *F-ratio* equal to $(n-2)\frac{\hat{\rho}^2}{1-\hat{\rho}^2}$; this one allow to do the test of hypothesis $\rho = 0$ against $\rho \neq 0$.

- finally, the last paragraph provides the $\hat{\rho}$ estimate of ρ, its square and the estimate of the standard deviation σ of the residuals. This number, *residual*, allows to build a confidence interval for σ.

Returning to the main screen allows you to choose the desired graphics. We can print:

- *plot fitted line*, the fitted curve,

- *plot residuals*, the residueal curve, which makes it possible to judge whether the mistakes are actually small or not,

- *save residuals* allows to put these errors in a file and to carry out tests on it,

- *save predictions,*

- *lack of fit test* .

The command **OUTLIER** allows to exclude points from the analysis (only for the linear model), for example: outliers, points of a past too far, points of the last period to use test points. The user manual is identical to that of REG.

2.3 Non linear regression

As we saw in the exercise Section 2.2.3, it is most often to reduce linear regression to model the least squares according to the functions listed below, for example:

$$\begin{aligned} f(t) \;=\; & a\cos(\omega t+\phi),\ \text{periodic function;}\\ & a\log t+b;\\ & \frac{1}{a+bt},\ \text{inverse function;}\\ & a.b^t;\, a.b^t+c,\ \text{exponent and modified exponent functions;}\\ & a.t^b;\, a.t^b+c,\ \text{power and modified power functions;}\\ & \frac{a}{1+b.c^t},\ \text{logistic function;}\\ & \exp[a.b^t+c],\ \text{Gompertz function.} \end{aligned}$$

If one can not simply reduce to the linear case, there are cases where one can minimize by the least squares method.
Exercise: parabola fitting.

This model is solved by the software using the NONLIN procedure where the formula of the adjustment function is explicitly given.

2.4 Multiple Regression

In some economic models, forecasting can be done using several explanatory variables. Indeed, prior statistical studies were able to detect "external" variables, particularly well correlated with the variable studied (to be explained, “internal”.) For example, a country’s energy consumption is a function of:

- industrial production,

- household consumption,

- number of cars, etc.

and we try to forecast by the multiple linear regression:

$$\hat{Y} = aX^1 + bX^2 + cX^3$$

where a, b, c are obtained by the least squares method, minimizing the application:

$$(a, b, c) \mapsto \sum_i (y_i - aX_1^i - bX_2^i - cX_3^i - d)^2.$$

Thus we get the linear system:

$$\begin{aligned}
\partial_a : a\sum_i (X_i^1)^2 + b\sum_i X_i^1 X_i^2 + c\sum_i X_i^1 X_i^3 + d\sum_i X_1^i &= \sum_i X_i^1 Y_i, \\
\partial_b : a\sum_i X_i^1 X_i^2 + b\sum_i (X_i^2)^2 + c\sum_i X_i^2 X_i^3 + d\sum_i X_2^i &= \sum_i X_i^2 Y_i \\
\partial_c : a\sum_i X_i^1 X_i^3 + b\sum_i X_i^3 X_i^2 + c\sum_i (X_i^3)^2 + d\sum_i X_3^i &= \sum_i X_i^3 Y_i \\
\partial_d : a\bar{X}_1 + b\bar{X}_2 + c\bar{X}_3 + d &= \bar{Y}.
\end{aligned}$$

Exercise : exhibit the optimal estimates $\hat{a}, \hat{b}, \hat{c}, \hat{d}$.

The routine **MREG** consists in obtaining the estimates of the coefficients a, b, c, d. We enter the name of the variable to be explained (or any other variable obtained by combination of what exists) it is *Dep var*, then the name of the explanatory variables is *Ind.var*.

The **STEP procedure** seeks, *stepwise*, the "best" variables to enter one by one, better in the sense that one firstly enter the best correlated with Y ($\rho(Y, X_i) = \sup_j \rho(Y, X_j)$), we operate the regression of Y with respect to X, then we choose the best correlated variable with the residual $Y - \hat{a}_i X_i$, and so on until the Fisher test on correlations becomes not significant.

Above has to be given in software R

2.5 Quality of the regression

In the linear case, the histogram of the residuals is examined to check that they are "acceptable." The program also provides some criteria:

ME (mean error) is the residuals mean, theoretically it could be 0,

MSE (mean square error) is the mean of the squared residuals, $s^2(\varepsilon)$,

MAE (mean absolute error) is the mean of absolute values of the residuals,

MAPE (mean absolute percentage error) is the mean of absolute values of the ratio residuals/their estimates; the interest is that this mean does not depend on the chosen unit,

MPE (mean percentage error) is the mean of the values of the ratio residuals/their estimates; once again, the interest is that this mean does not depend on the chosen unit,

2.6 Durbin Algorithm

Perhaps Section to skip.
This method "mixes" both regression and smoothing: indeed, the aim is to forecast X_{n+1} using the k previous observations, meaning $(X_n, X_{n-1}, \cdots, X_{n-k+1})$ using the least square method:

$$\hat{X}_{n+1}(k) = \sum_{i=1}^{k} a_i(k) X_{n+1-i} \tag{12}$$

is the L^2 projection on the vectorail subspace generated by $(X_n, X_{n-1}, \cdots, X_{n-k+1})$, so the parameters $a_i(k)$ realize the minimum of the application:

$$F : (a_i(k), i = 1, \cdots, k) \to E[X_{n+1} - \sum_{i=1}^{k} a_i(k) X_{n+1-i}]^2$$

the minimum (=the error) is denoted as v_k.

Thus we need some assumptions on the process X_n :
Hypothesis : the observations x_i are the observed values of the random variables X_i, centered, square integrable, stationary, meaning: for all n , $cov(X_n, X_{n+k}) = \gamma(k)$. Remark that the function γ is pair on Z: $\gamma(k) = \gamma(-k)$.
Definition : the function γ is the **autocovariance function** .

This function could be estimated using the observed values. Then

Proposition 2.8. *Under the previous assumptions, we get the recursive relations:*

$$a_1(1) = \frac{\gamma(1)}{\gamma(0)} \; ; \; a_i(k) = a_i(k-1) - a_k(k) a_{k-i}(k-1) \; ; \; i = 1, ...k-1, \tag{13}$$

$$v_0 = \gamma(0) \; ; \; v(k) = v_{k-1}(1 - a_k(k)^2), k \geq 1 \tag{14}$$

$$a_k(k) = \frac{\gamma(k) - \sum_{i=1}^{k-1} \gamma(k-i) a_i(k-1)}{v_{k-1}}, k \geq 2. \tag{15}$$

Proof:

(i) $k = 1$: $\hat{X}_2 = a_1(1) X_1$ where $a_1(1)$ realizes the minimum of the application $a_1 \to E[(X_2 - a_1 X_1)^2]$. This application is a convex differentiable function, so its minimum is realized when the derivative is null: $a_1(1) = \frac{\gamma(1)}{\gamma(0)}$.

Let us denote:

$$E_{n,k} = \text{ the vectorial subspace generated by } (X_n, \cdots, X_{n-k+1}).$$

and $P_{n,k}$ the projector on $E_{n,k}$. Obviously

$$\hat{X}_{n+1}(k) = P_{n,k}(X_{n+1}) = \sum_{i=1}^{k} a_i(k) X_{n+1-i}.$$

We project this equality on the smaller vector space $E_{n,k-1}$ thus on the one hand

$$P_{n,k-1}(\hat{X}_{n+1}) = P_{n,k-1}(X_{n+1}) = \sum_{i=1}^{k-1} a_i(k-1) X_{n+1-i},$$

and on the other hand

$$P_{n,k-1}(\hat{X}_{n+1}) = \sum_{i=1}^{k-1} a_i(k) X_{n+1-i} + a_k(k) P_{n,k-1}(X_{n-k+1}).$$

We then use the lemma:

Lemma 2.9. *There is a "symmetry" between past and future, meaning:* $P_{n+k,k}(X_n) = \sum_{i=1}^{k} a_i(k) X_{n+i}$.

Proof: we get (12) by minimizing the application F, differentiable convex function, thus $a_i(k)$ are solution of the linear system

$$\nabla^i F = -2\gamma(i) + 2\sum_{j=1}^{k} a_j \gamma(|i-j|) = 0, i = 1, \cdots, k.$$

In the lemma, X_{n+i} coefficients have to minimize the application
$a_i \to G(a_i) = E[X_n - \sum_{i=1}^{k} a_i(k) X_{n+i}]^2$. We can check that actually both linear systems are the same; this ends the proof. •

Thus

$$P_{n,k-1}(X_{n-k+1}) = a_1(k-1) X_{n-k+2} + a_2(k-1) X_{n-k+3} + \cdots + a_{k-1}(k-1) X_n.$$

We identify both expressions of $P_{n,k-1}(\hat{X}_{n+1})$, then yields the coefficient of X_{n+1-i} under two expressions:

$$a_i(k-1) = a_i(k) + a_k(k) a_{k-i}(k-1),$$

meaning (13).

(ii) The projection on $\{0\}$ is necessarily null, $\hat{X}_{n+1} = 0,\ v_0 = \| X_{n+1} \|^2 = \gamma(0)$.

For computing the error $v_k$we use the Pythagore Theorem:

$$v_k = |X_{n+1} - P_{n,k}(X_{n+1})|_2^2,$$

so:

$$v_{k-1} = |X_{n+1} - P_{n,k-1}(X_{n+1})|_2^2 = v_k + |P_{n,k}(X_{n+1}) - P_{n,k-1}(X_{n+1})|_2^2.$$

But

$$P_{n,k}(X_{n+1}) - P_{n,k-1}(X_{n+1}) = a_k(k)[X_{n-k+1} - P_{n,k-1}(X_{n-k+1})].$$

Using Lemma 2.9 and the stationarity of the process X, the squared norm of the vector $P_{n,k}(X_{n+1}) - P_{n,k-1}(X_{n+1})$ is $a_k(k)^2 v_{k-1}$, so:

$$v_{k-1} = v_k + a_k(k)^2 v_{k-1}$$

meaning (14).

(iii) Using once again

$$P_{n,k}(X_{n+1}) - P_{n,k-1}(X_{n+1}) = a_k(k)[X_{n-k+1} - P_{n,k-1}(X_{n-k+1})]$$

we compute $v_{k-1} = |X_{n-k+1} - P_{n,k-1}(X_{n-k+1})|^2$. Actually, using both forms of this vector yields:

$$\langle X_{n-k+1} - P_{n,k-1}(X_{n-k+1}), P_{n,k}(X_{n+1}) - P_{n,k-1}(X_{n+1}) \rangle = a_k(k) v_{k-1}$$

Using the definition of the projector $P_{n,k-1}(X_{n+1}) \in E_{n,k-1}$ and the left factor above is orthogonal to $E_{n,k-1}$:

$$a_k(k)v_{k-1} = \langle X_{n-k+1} - P_{n,k-1}(X_{n-k+1}), P_{n,k}(X_{n+1})\rangle = \langle X_{n-k+1} - P_{n,k-1}(X_{n-k+1}), X_{n+1}\rangle.$$

But $P_{n,k-1}(X_{n-k+1}) = \sum_{i=1}^{k-1} a_i(k-1)X_{n+1-i}$ so:

$$a_k(k)v_{k-1} = \gamma(k) - \sum_{i=1}^{k-1} a_i(k-1)\gamma(k-i).$$

•

Remark : One can interpret $a_k(k)$ as following: this coefficient is equal to the correlation coefficient between the vectors $X_{n+1} - P_{n,k-1}(X_{n+1})$ and $X_{n-k+1} - P_{n,k-1}(X_{n-k+1})$. It is named **partial correlation coefficient**. Indeed, remind the proof of (iii):

$$a_k(k) = \frac{\langle X_{n-k+1} - P_{n,k-1}(X_{n-k+1}), X_{n+1}\rangle}{v_{k-1}} \tag{16}$$

$$= \frac{\langle X_{n-k+1} - P_{n,k-1}(X_{n-k+1}), X_{n+1} - P_{n,k-1}(X_{n+1})\rangle}{v_{k-1}} \tag{17}$$

so v_{k-1} is actually the squared norm of these vectors.

2.7 Innovation Algorithm

(cf [6], page 155 et sq.)

For forecasting, it is interesting not to recalculate all the coefficients each time a new information arrives while using it nevertheless. We therefore try to use the estimates already obtained as well as the new observation to predict at best.

Definition 2.10. Innovation *is the "new" information at time t, meaning:*

$$Z_t = X_t - P_{t-1}(X_t),$$

where P_t is the projection on E_t, vectorial subspace generated by $\{X_1, \cdots, X_t\}$.

Remarks :
(i) E_t is also generated by the vectors $\{Z_1, \cdots, Z_t\}$.
(ii) By construction, the vectors Z_i are mutually orthogonal vectors.

Let v_{t-1} denote $E(X_t - P_{t-1}(X_t))^2$, the quadratic error or "risk". Particularly $v_0 = E(X_1 - P_0(X_1))^2 = E(X_1)^2 = \gamma(0)$ since $P_0(X_1) = 0$.

Proposition 2.11. *Let $P_t(X_{t+1})$ denote $\sum_{j=1}^{t} \tau_j(t)Z_j$, then recursively:*

$$\tau_1(t) = \frac{\gamma(t)}{\gamma(0)}$$

$$\tau_j(t) = \frac{1}{v_{j-1}}[\gamma(t+1-j) - \sum_{i=1}^{j-1} \tau_i(j-1)\tau_i(t)v_{i-1}];\ j \geq 2 \tag{18}$$

$$v_0 = \gamma(0), v_t = \gamma(0) - \sum_{j=1}^{t} \tau_j^2(t)v_{j-1}\ ;\ t \geq 1 \tag{19}$$

Proof: The recursion starts with $P_0(X_1) = 0$ so $Z_1 = X_1$ and $v_0 = \gamma(0)$.

Firstly remark that for $j = 1$ minimizing the application $a \Rightarrow E(X_2 - aX_1)^2$ yields $\tau_1(1) = \frac{\gamma(1)}{\gamma(0)}$.

Since the Z_j are orthogonal, if $j \leq t$:

$$\langle X_{t+1}, Z_j \rangle = \langle P_t(X_{t+1}), Z_j \rangle = \tau_j(t) E(Z_j^2) = \tau_j(t) v_{j-1}.$$

This means that

$$\tau_j(t) = \frac{\langle X_{t+1}, X_j - P_{j-1}(X_j) \rangle}{v_{j-1}};$$

writing $P_{j-1}(X_j) = \sum_{i=1}^{j-1} \tau_i(j-1) Z_i, j \geq 2$, we have to compute the scalars products $\langle X_{t+1}, Z_i \rangle = \langle P_t(X_{t+1}), Z_i \rangle = \tau_i(t) v_{i-1}$, and for $j = 1, \tau_1(t) = \frac{\gamma(t)}{\gamma(0)}$.
Thus (18) is proved.
To prove (19), once again we use the Pythagore theorem:

$$E(X_{t+1})^2 = v_t + E[P_t(X_{t+1})]^2,$$

this is exactly (19) since $P_t(X_{t+1})$ is a sum of mutually orthogonal vectors. •

For further forecasting, we have the following proposition :

Proposition 2.12.

$$\hat{X}_{n+k}(n) = P_n(X_{n+k}) = \sum_{i=1}^{n} \tau_i(n+k-1) Z_i.$$

Proof: standard, with

$$P_n(X_{n+k}) = P_n \circ P_{n+k-1}(X_{n+k}) = P_n[\sum_{i=1}^{n+k-1} \tau_i(n+k-1) Z_i].$$

Since Z_i are mutually orthogonal vectors, the projection of $Z_i, i > n$ on E_n is null, so the result. •

Practically, the rule is to replace by 0 the innovations of instants $n+1$ to $n+k-1$ which actually are still unknown.

3 Box and Jenkins' methods, general features

Developed in the 70s, these are very powerful methods which make maximum use of the fact that the evolution of the studied time series is considered as **one** of the achievements of a stochastic process, endowed with a strong enough structure. Indeed, once highlighted the structure, this allows to predict more confidently the future series. The counterpart is the need for a fairly long period of observations for that the forecast is reliable. The authors recommend 5 **to** 6 **periods** in the case of periodic phenomena, and a minimum of 30 observations in other cases.

These methods work very well for short-term forecasts macroeconomic series, especially for the industrial production indexes. In Finance, this method does not concern the forecast of returns, but the one of volatility.

They are based on the assumption that each observation depends quite strongly on previous observations. Basically, this addiction to the past replaces multiplicity of observations (in Statistics) to estimate the settings by applying the law of large numbers. So are assumed strong enough assumptions, that the series is stationary, meaning the two first moments do not depend on time. If this is not the case, they must be done "stationary" by transformations (called filters) that remove trend and seasonality.

3.1 Definitions

Thus, we consider processes, random series, indexed in $\mathbb{Z}$ and taking their values in $\mathbb{R}$ (real numbers):

$$\forall n \in \mathbb{Z},\ X_n \text{ is a random variable } : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B}).$$

We try to model the application $n \mapsto X_n$ with a trend part, a seasonal component, and the measurement error.

Hypothesis: The observations x_n are the values of a centered, square-integrable, stationary, random process (X_n), i.e. there exists a function γ on $\mathbb{Z}$ such that $\forall n, k \in \mathbb{Z}$, $cov(X_n, X_{n-k}) = \gamma(k)$, where

$$cov(X_n, X_{n-k}) = E[(X_n - E(X_n))(X_{n-k} - E(X_{n-k}))].$$

Exercise 1: Actually for any $k \in \mathbb{Z}$, $\gamma(k) = \gamma(-k)$.

Definition 3.1. *: Such a process is called a* **second order stationary time series**, *S.T.S. for short.*

The function γ is called the **auto covariance function**.

Moreover we define the **auto correlation function** $\rho : k \mapsto \rho(k) = \frac{\gamma(k)}{\gamma(0)}$.

There exists another notion: "strict stationarity" meaning the vectors $(X_1, \cdots, X_k)$ and $(X_{n+1}, \cdots, X_{n+k})$ have the same law, for any pair (k, n).

As for the covariance function γ, for any $k \in \mathbb{Z}$, $\rho(k) = \rho(-k)$ and we define the **correlogram**, graph of the application ρ, useful tool for analyzing the series as discussed later.

We also introduce:

Definition 3.2. *The* **partial auto correlation function**, *P.A.C.F., is defined on* $\mathbb{N}$ *as:*

$$r : \mathbb{N} \to \mathbb{R} \; ; \; r(p-n) = \; cor\,(X_n, X_p / X_{n+1}, \cdots, X_{p-1}), p > n,$$

meaning

$$r(p-n) = \frac{cov\,(X_n - X_n^*, X_p - X_p^*)}{\sqrt{Var\,(X_n - X_n^*)\;\; Var\,(X_p - X_p^*)}}$$

where X_j^* *is the orthogonal projection of* X_j *on the vector space* $S_{n,p}$ *generated by* $(X_{n+1}, \cdots, X_{p-1})$, *and completed by* $r(1) = \rho(1)$.

Exercise 2: this expression only depends on $(p-n)$.
Finally, we introduce the infinite dimensional matrix of variance-covariance process X.

Definition 3.3. *: The* **Toeplitz matrix** *is*

$$\Gamma, \;\; \gamma(i,j) = r(i-j), \;\; i,j \geq 1.$$

This is a symmetric matrix.

3.2 Examples of second order stationary times series, STS

First example of fundamental S.T.S. : the white noise.

Definition 3.4. *The (weak)* **white noise** *is a STS* (ε_k) *(with covariance function equal to* γ *with* $\gamma(k) = \sigma^2 \delta_{k,0}$.

If moreover there is independence between the random variables (ε_k), the white noise is said **strong**.
For example, it could be a Gaussian process with covariance matrix $\Gamma = \sigma^2 I_d$; in this case, there is in addition the orthogonality of the white noise components ε_n in L^2 and their independence, thanks to the Gaussian nature of the series.
A *strong* white noise is a white noise such that (ε_n) are i.i.d. (independent identically distributed).
This "white noise process" is used to model the measurement error. If the series is not centered, the term is named "colored noise".

Second example:

Definition 3.5. *A* **moving average** *is a STS as follows:*

$$X_n = \sum_{k\in\mathbb{Z}} a_k \varepsilon_{n-k},$$

where the series $(a_k; k \in \mathbb{Z}) \in l^2$ *and* ε *is a white noise.*
For short: M.A.= "moving average".

Proposition 3.6. *The covariance function of a moving average* $X_n = \sum_{k\in Z} a_k \varepsilon_{n-k}$ *is written as* $\gamma(p) = \sum_{k\in Z} a_{p-k} a_{-k}$ $\forall p \in \mathbb{Z}$.

Proof: : We write X_n and X_{n-p} definition; firstly remark that these series are L^2 convergent using the hypothesis that the series $(a_k; k \in Z) \in l^2$. Secondly we compute their covariance, meaning the mean of the product since these random variables are centered:

$$E[X_n X_{n+p}] = \lim_{K\to\infty} \sum_{|k|<K} a_{p+k} a_k.$$

This limit exists since

$$\forall K > 0, \quad (\sum_{|k|<K} a_{p+k} a_k)^2 \leq \sum_{|k|<K} |a_{p+k}|^2 \sum_{|k|<K} |a_k|^2 < \infty.$$

This inequality is proved recursively: it is true for $K = 2$*, and the property for* $K - 1$ *implies it for* K. •

Definition 3.7. *When there exists a finite number of non null coefficients* a_k*, i.e.* $(a_0, \cdots, a_p)$*, we say that* X *is a* **order** p**-moving average**, *MA(p) for short.*

Third example: let ε be a white noise, and define the recursive series

$$X_n = \alpha X_{n-1} + \varepsilon_n.$$

Assuming that we know a particular element of the series, for instance X_0, assuming it is a centered random variable in L^2 we prove the following.

Proposition 3.8. *Let* X *be the process defined as*

$$X_n = \alpha X_{n-1} + \varepsilon_n, \ \forall n \in \mathbb{Z}, \ X_0 \in L^2, E[X_0] = 0.$$

Assuming $|\alpha| < 1$*, and* $E[X_n^2] \leq M^2$, $\forall n \in \mathbb{Z}^-$*, then* X *is a STS.*

Specifically, this is a moving average with coefficients $a_j = \alpha^j$, $j \geq 0$*. Its covariance function is defined by* $\gamma(k) = \frac{\alpha^k}{1-\alpha^2}$.

Exercise: prove this result.

Definition 3.9. *An order* 1 *auto regressive series* X *(AR*(1) *for short) is a process depending only of the previous observation, step by step.*

At this point we can quote Francq and Zakoian [4] pp 7-11: Sections 1.3 *Financial Series* **and Section 1.4** *Random variance models* **which show how ARMA processes are not appropriate to model Financial Series as it is written above in the introduction**

Indeed, once again, the financial data present some stylised facts:

- non stationarity of price series,
- absence of auto correlation for the price variations,
- unpredictability of returns,
- auto correlation of the squared price returns,
- volatility clustering $\Rightarrow$ prediction of squared returns,
- fat tailed distributions (leptokurticity),
- leverage effects,
- seasonality.

3.3 Delay Operator, ARMA equations

In this subsection we consider that X is a STS. In AR(1) example, $X_n = aX_{n-1} + \varepsilon_n$ and $\forall(\varepsilon_n)$ (a given white noise) we get X_n as a function of X_{n-1}; more generally it is interesting to get formal this passage from $n-1$ to n. Firstly we have to define the spaces on which is defined this passage.

Definition 3.10. *The closed subspace generated by the set* $\{X_p, p \in \mathbb{Z}, p \leq n\}$ *in* L^2 *is denoted as* H_n^X.
This subspace of L^2, H_n^X, *is named the* **linear past** *of* X.
We note also:

$$H_{-\infty}^X = \cap_n H_n^X \ ; \ H_{+\infty}^X = \cup_n H_n^X = H^X.$$

$H_{-\infty}^X$ *is named the asymptotic past,* H^X *the linear envelope.*

These spaces are used to characterize two specific types of STS.
Following Francq and Zakoian [4] page 4, we consider $\varepsilon_{\mathbf{n}} := \mathbf{X_n} - \mathbf{P_{n-1}}(\mathbf{X_n})$, weak or strong white noise, where P_{n-1} is the L^2 orthogonal projector on H_{n-1}^X.

Definition 3.11. *When* $H_{-\infty}^X = \{0\}$ *the series is* **regular**.
When $H_{-\infty}^X = H^X$ *the series is* **singular**. *In this case, the linear pasts are constant and the "innovation" does not bring any information.*

A first example of regular STS is the white noise: Actually because the process ε is non correlated, the vector space $H_n^\varepsilon = \mathbb{R}\varepsilon_n + H_{n-1}^\varepsilon$. So if $Y \in H_n^\varepsilon \cap H_{n-1}^\varepsilon$, firstly, $Y = a\varepsilon_n + P_{n-1}^\varepsilon(Y)$. But $Y \in H_{-\infty}^\varepsilon$ means that $Y \in H_{n-1}^\varepsilon$, so $a = 0$. And so on, $Y = 0$ and ε is a regular series.

Exercise, other examples: Look at the regularity of the following SCS:

$X_n = g(n)X_0$ where g is an application from $\mathbb{Z}$ to $\mathbb{R}$ such that X is a SCS.

A white noise, a moving average, a unilatere moving average, an AR(1).

Definition 3.12. *The operator $H^X = vect\{X_n,\ n \in Z\}$ in L^2 which maps X_n to X_{n-1} is named the* **delay operator** *denoted S^X : $S^X(X_n) = X_{n-1}$.*

Proposition 3.13. *The operator S^X is the unique isometry from H^X to H^X which sends X_n to X_{n-1}. Moreover, $S^X(H^X) = H^X$.*

Proof: : The operator S^X is defined on the $\{X_n, n \geq 0\}$ and is extended by linearity on any finite linear combinations of X_n. This is an isometry:

$$\begin{aligned} \| S^X(\sum_i a_i X_i) \|_2^2 &= \sum_{i,j} a_i a_j E[X_{i-1}X_{j-1}] \\ &= \sum_{i,j} a_i a_j \gamma(i-j) = \| \sum_i a_i X_i \|_2^2 . \end{aligned}$$

Thus we could extend this operator S^X by continuity on the whole H^X.
Uniqueness: it is a consequence of the fact that if T could be another solution, $T = S_X$ on any X_n, so on any finite linear combinations of X_n so by continuity on H^X.
Any element of H^X is a limit of finite linear combinations of X_n, image by S^X of finite linear combinations of X_n, so the equality $S^X(H^X) = H^X$. •

Theorem 3.14. *(WOLD): Any STS could be written as a unique sum of a regular and a singular parts:*

$$X = X^r + X^s$$

so that the spaces H^{X^r} and H^{X^s} are L^2 orthogonal.

Proof: Exercise, using $X_n^s := P_{-\infty}^X(X_n)$; $X_n^r := X_n - P_{-\infty}^X(X_n)$.
(i) By definition, $X_n = X_n^s + X_n^r$,
(ii) Any $Y \in H_{-\infty}^{X^r} \subset H_{-\infty}^X$, Y is orthogonal to $H_{-\infty}^X$ and $Y \in H_{-\infty}^X$ .... so $Y = 0$ and X^r is a regular series.
(iii)Let $Y \in H_n^{X^s}$ for any n, by definition of X^s there exists $Z_n \in H_n^X$ such that $Y = P_{-\infty}^X(Z_n)$. Thus $Y \in H_{-\infty}^X$ and for any n $H_n^{X^s} \subset H_{-\infty}^X$.
Conversely, let $Y \in H_{-\infty}^X$. So on the one hand $Y = P_{-\infty}^X(Y)$ and on the other hand $Y \in H_n^X$ for any n could be written as $Y = \sum_{n_i \leq n} a_{n_i} X_{n_i} = \sum_{n_i \leq n} a_{n_i} P_{-\infty}^X X_{n_i} = \sum_{n_i \leq n} a_{n_i} X_{n_i}^s \in H_n^{X^s}$.

•

Proposition 3.15. *Both series X^r and X^s are too STS.*

Proof: : Firstly by construction they are centered and in L^2.
Secondly we use the following:

Lemma 3.16. *For all $n \in Z$, $P_n^X \circ S^X = S^X \circ P_{n+1}^X$.*

Proof: for all $p \in Z$, $P_n^X \circ S^X(X_p) = P_n^X(X_{p-1})$ is the unique vector in H_n^X such that $X_{p-1} - P_n^X(X_{p-1})$ is orthogonal to H_n^X. So we have to compute $\forall k \leq n$ the scalar product $\langle X_k, X_{p-1} - S^X \circ P_{n+1}^X(X_p)\rangle$. This scalar product is equal to:

$$\begin{aligned}\langle X_k, X_{p-1} - S^X \circ P_{n+1}^X(X_p)\rangle &= \gamma(k-p+1) - \langle S^X(X_{k+1}), S^X \circ P_{n+1}^X(X_p)\rangle \\ &= \gamma(k-p+1) - \langle X_{k+1}, P_{n+1}^X(X_p)\rangle\end{aligned}$$

since S^X is an isometry. Then we use $\forall k \leq n, X_{k+1} \in H_{n+1}^X$. Yields:

$$\langle X_k, X_{p-1} - S^X \circ P_{n+1}^X(X_p)\rangle = \gamma(k-p+1) - \langle X_{k+1}, P_{n+1}^X(X_p)\rangle = \gamma(k-p+1) - \langle X_{k+1}, X_p\rangle = 0.$$

•

We apply this lemma to the computation of the covariance function of the series X^s, with $n \geq p$:

$$\begin{aligned}(X_n^s, X_p^s) = (P_{-\infty}^X(X_n), P_{-\infty}^X(X_p)) &= (S^X \circ P_{-\infty}^X(X_n), S^X \circ P_{-\infty}^X(X_p)) = \\ (P_{-\infty}^X \circ S^X(X_n), P_{-\infty}^X \circ S^X(X_p)) &= (P_{-\infty}^X(X_{n-1}), P_{-\infty}^X(X_{p-1}))\end{aligned}$$

which is exactly (X_{n-1}^s, X_{p-1}^s) by definition of X^s, step by step we go to

$$(X_n^s, X_p^s) = (X_{n-p}^s, X_0^s),$$

which only depends on the difference $n - p$; this proves the stationarity of the series (X^s). Then, the part $X^r = X - X^s$ is too a STS: $X^r \in L^2$ with null expectation by linearity, and we easily check the stationarity of $E[(X_n^r, X_p^r)]$. More specifically *using $E(X_n X_p^s) = \gamma^s(n-p)$* we prove:

$$(X_n - X_n^s, X_p - X_p^s) = \gamma(n-p) - \gamma^s(n-p).$$

This shows the stationarity of X^r and the relation between the covariance functions $\gamma = \gamma^r + \gamma^s$. •

Remark 3.17. *When a STS is not singular, the strict inclusion $\forall n$, $H_{n-1}^X \subset H_n^X$ is satisfied. Indeed, if not, there exists n such that $H_{n-1}^X = H_n^X$, and with the lemma and the delay operator S^X we deduce that $\forall n$, $H_{n-1}^X = H_n^X$, so the series is singular.*

The following theorem provides a characterization of regular series.

Theorem 3.18. *A series X is regular if and only if there exists a series (d_n) in $l^2(\mathbb{R})$ and a white noise ε such that:*

$$X_n = \sum_{p \geq 0} d_p \varepsilon_{n-p}.$$

We could choose ε so that the linear pasts of X and ε are identical; then this white noise and the associated series (d_n) are unique, except a possible multiplicative coefficient.

Indication for the proof: recursively, define $\varepsilon_n = X_n - P^X_{n-1}X_n$, and

$$X_n = \sum_{j=0}^{p+1} d_j \varepsilon_{n-j} + P^X_{n-p}X_n \quad \textit{and let } p \to \infty.$$

Definition 3.19. *This specific white noise is named* **innovation white noise**.

The interest of such series lies in the following corollary: the projection on the past is then extremely simple.

Corollary 3.20. *Let X be a regular series and ε its innovation white noise; for all $m \leq n$,*

$$P^X_m(X_n) = \sum_{p \geq n-m} d_p \varepsilon_{n-p}.$$

Proof: of the theorem:
By definition $X_n \in H^\varepsilon_n$, so $H^X_n \subset H^\varepsilon_n$, $\cap_n H^X_n \subset \cap H^\varepsilon_n = \{0\}$ since ε is regular, and X is regular.

Conversely, let X be a regular series. Let the process $v_n = X_n - P^X_{n-1}(X_n)$; this is a STS since we could compute its covariance function:

$$\forall n, \quad \| v_n \| = \| S^X(X_{n+1}) - P^X_{n-1} \circ S^X(X_{n+1}) \| = \| X_{n+1} - P^X_n(X_{n+1}) \| = \| v_{n+1} \|$$

denoted $\sigma^2 = \gamma(0)$. By definition, $v_n \in H^X_n$ and is orthogonal to H^X_{n-1} so to the previous v_i: thus it is a STS, and more specifically it is a white noise denoted $a_0\varepsilon_n$.

By definition, $X_n = a\varepsilon_n + P^X_{n-1}(X_n)$, $\varepsilon_n \in H^X_n$ and is orthogonal to H^X_{n-1}, thus H^X_n is the direct sum $\mathbb{R}\varepsilon_n \oplus H^X_{n-1}$. By induction we get that H^X_n is the direct sum $\oplus_{0 \leq i \leq j} \mathbb{R}\varepsilon_{n-i} \oplus H^X_{n-j-1}$. On this direct sum we get the decomposition

$$X_n = \sum_{0 \leq i \leq j} a_i \varepsilon_{n-i} + P^X_{n-j-1}(X_n)$$

Since X is a regular series, $\lim_{j \to \infty} P^X_{n-j-1}(X_n) = 0$ and X is equal to $\sum_{0 \leq i} a_i \varepsilon_{n-i}$, which is the expected form.

As a consequence, $X_n \in H^\varepsilon_n$ and since previously we knew that, $\varepsilon_n \in H^X_n$, these two spaces are identical.

Uniqueness: we assume that there exists a pair (ε', d'), (white noise, $l^2(C)$ element), solution of the problem, so satisfying

$$\forall n, P^{\varepsilon'}_n = P^X_n = P^\varepsilon_n \text{ and } X_n = \sum_{0 \leq i} d_i \varepsilon_{n-i} = \sum_{0 \leq i} d'_i \varepsilon'_{n-i}.$$

On both hands of this equality we apply the operator P^X_{n-1}, we get :

$$P^\varepsilon_{n-1}(X_n) = \sum_{1 \leq i} d_i \varepsilon_{n-i}; P^{\varepsilon'}_{n-1}(X_n) = \sum_{1 \leq i} d'_i \varepsilon'_{n-i}.$$

But $P_n^{\varepsilon'} = P_n^{\varepsilon}$ so the difference is null and $\forall n$, $d'_0\varepsilon'_n = d_0\varepsilon_n$ meaning the uniqueness except a possible multiplicative coefficient. •

The proof of the corollary is obvious since the operators P_m^X and P_m^ε are the same, as are the corresponding spaces H_m^X and H_m^ε.

Remark 3.21. *The identity between these two families of spaces is interpreted as follows: Linear pasts of X and ε coincide. If X is observed up to time $n-1$, the additional information provided by really new X_n is represented by $a\varepsilon_n = X_n - P_{n-1}^X(X_n)$, the 'innovation' as we called it previously.*

More generally, we will now study the class of STS, solution of " ARMA " equations, written using the delay operator S^X.

Definition 3.22. *Let Xbe a STS and let ε be a white noise, P and Q two polynomials. We say that X* **is solution of ARMA(P,Q) equation** *if this process satisfies for any n in $\mathbb{Z}$:*

$$P(S^X)(X_n) = Q(S^\varepsilon)(\varepsilon_n), \tag{20}$$

meaning there exist complex coefficients $(a_0, \cdots, a_p)$ and $(b_0, \cdots, b_q)$ such that $\forall n \in \mathbb{Z}$:

$$\sum_{i=0}^{p} a_i X_{n-i} = \sum_{i=0}^{q} b_i \varepsilon_{n-i}. \tag{21}$$

In case of $p = 0$, X is **MA(q)** ; in case of $q = 0$, X is **AR(p)**. In the general case we say that X is **ARMA(p,q)**.
Such an equation could be solved, either to get X function of process ε or the converse so that we could "forecast" X_n based solely on its past. Roughly speaking, this consists in a "reverse" of operators $P(S^X)$ and $= Q(S^\varepsilon)$. This is out of our agenda, but the following Section 3.4 is an important result which will be useful in the second part of this course.

3.4 ARMA Equation: resolution

Let $A_P(X) = A_Q(\varepsilon)$ an ARMA equation.

Theorem 3.23. *(Fejer-Riesz) Let P et Q be non nul polynomials with no common roots, those of P have modulus $\neq 1$. Then the ARMA equation is solvable as soon as the modulus of P roots are > 1 and those of $Q \geq 1$.*

Definition 3.24. *This ARMA representation of X is called* **canonical Fejer-Riesz canonical representation**.

3.5 Estimate of the covariance function of an ARMA Process

We come back to the observation of a STS, supposed to be stationary, non necessarily centered:

$$X_1, \cdots, X_N,$$

The first step is to estimate $E(X)$ and the covariance function γ_X.

According to standard probability or statistics lecture notes in case of sampling, $E(X)$ is estimated by Cesàro mean, that is justified by the large numbers law (cf. [1]):

$$\widehat{E(X)} \sim \frac{1}{n}\sum_{i=1}^{n} X_i.$$

But the required assumptions are either the independence of the observations or the martingale property for the process. Neither of these assumptions is checked in the case of STS. Nevertheless, with similar proofs to those found in a Probability course, we get same type results. This is what will be used to justify approximates of the mean and of the covariance function.

Insert work with R: 'plotobs(X)' to draw the series graph; mean(X); acf(X) to get correlogram, variogram, partial correlogram...see TD-TP Agnes Lagnoux.

3.6 Large numbers law

Lemma 3.25. *Let $X_1, \cdots, X_n,\ n \in N$ be a series of random variables with mean m. We put $S_n := \sum_{i=1}^{n} X_i$ and assume:*

$$\exists M > 0,\ Var(X_n) \leq M^2,\ Var(S_n) \leq nM^2,\ \forall n \geq 1.$$

Then $\frac{1}{n}S_n \to m$ in L^2 and almost surely, when n goes to infinity.

Proof: : Exercise.
(i) $Var(\frac{1}{n}S_n) = E[\frac{1}{n}S_n - m]^2$ since by hypothesis $E(S_n) = nm$. But $Var(\frac{1}{n}S_n) = \frac{1}{n^2}Var(S_n) \leq \frac{1}{n}M^2 \to 0$ when n goes to infinity, so yields the convergence in L^2.

(ii) Let $Z_k = \sup\{|\frac{1}{n}S_n - m|,\ n \in [k^2, (k+1)^2[\}$. We put $Y_j := X_j - m$ so:

$$\frac{1}{n}S_n - m = \frac{1}{n}S_{k^2} + \frac{1}{n}(X_{k^2+1} + \cdots + X_n - nm) = \frac{1}{n}(S_{k^2} - k^2 m + Y_{k^2+1} + \cdots + Y_n).$$

Then we deduce the bound:

$$Z_k \leq \frac{1}{k^2}(|S_{k^2} - k^2 m| + |Y_{k^2+1}| + \cdots + |Y_{(k+1)^2-1}|)$$

so the L^2 norm satisfies:

$$\|Z_k\|_2 \leq \frac{1}{k^2}(\|S_{k^2} - k^2 m\|_2 + \|Y_{k^2+1}\|_2 + \cdots + \|Y_{(k+1)^2-1}\|_2).$$

By hypothesis the first term is bounded by Mk, and any of the $(k+1)^2 - 1 - k^2 = 2k$ following terms are equal to the X standard deviation bounded by M:

$$\|Z_k\|_2 \leq \frac{1}{k^2}(Mk + 2kM) = 3M/k.$$

Thus the series $E(\sum_k Z_k^2) = \sum_k E(Z_k^2) \leq \sum_k 9M^2/k^2$ is convergent, proving that Z_k converges almost surely, when k goes to infinity, exactly meaning $\frac{1}{n}S_n - m$ converges almost surely to zero, meaning $\frac{1}{n}S_n$ converges almost surely to m when n goes to infinity. •

We apply this lemma to a STS: since $Var(X_n) = \gamma^X(0)$ the first hypothesis is satisfied. The second hypothesis concerns

$$Var(S_n) = Var(\sum_{i=1}^{n} X_i) = \sum_{1 \leq i,j \leq n} \gamma^X(i-j) = n\gamma^X(0) + 2(n-1)\gamma^X(1) + \cdots + 2\gamma^X(n-1)$$

the bound of which not necessarily being nM.
But for instance a MA(q) process satisfies this hypothesis since in this case there exists a finite number of non null $\gamma^X(i)$, $\gamma(k) = 0$ for all $k > q$:

$$Var(S_n) \leq n(\gamma(0) + \cdots + \gamma(q)).$$

Exercise: under the assumption of the lemma above, in case of an AR(1), $X_n = aX_{n-1} + \varepsilon_n$ prove that the covariance is $\gamma^X(k) = \frac{a^k}{1-a^2}$.

3.7 Covariance function estimate, *acf*, *pacf*

Let k be fixed in $\mathbb{N}$ (if $k < 0$, $\gamma(k) = \gamma(-k)$). Using the large numbers law (or rather Lemma 3.25), if the series $Y : n \to X_n X_{n+k}$ has "good" properties, a $\gamma^X(k)$ reasonable estimate is:

$$\bar{\gamma}_n(k) = \frac{1}{n}\sum_{j=1}^{n} X_j X_{j+k}.$$

For that remark that we need observations at least from time 1 to $n+k$.
If we have only n observations, we propose:

$$\gamma_n^*(k) = \frac{1}{n}\sum_{j=1}^{n-k} X_j X_{j+k}.$$

Both estimates have the following properties:

(i) **Bias**

$$E[\bar{\gamma}_n(k)] = \gamma(k),$$

meaning this estimate has a null bias $\forall n$.

$$E[\gamma_n^*(k)] = \frac{n-k}{n}\gamma(k) \to \gamma(k),$$

this estimate bias is asymptotically null.
Exercise: compute the bias of these both estimates.

(ii) **Convergence and quadratic error**: here we need more hypotheses. To apply Lemma 3.25, $E(X_nX_{n+k}) = \gamma(k)$ but we also need the existence of a constant M such that $Var(X_nX_{n+k}) \leq M^2$ and $Var(\sum_{i=1}^n X_iX_{i+k}) \leq nM^2$ meaning we would need at least $X \in L^4$ and $\sup_n E(X_n^4) \leq M^2$. Now we detail the second hypothesis:

$$\sum_{1\leq i,j\leq n} E[X_iX_{i+k}X_jX_{j+k}] - n^2\gamma^2(k) \leq nM^2$$

we could (for instance) assume that the series distribution is Gaussian.

Be careful: in case of financial series, it is a stylized fact that price processes are not Gaussian, thus in such a case we can not use this hypothesis.
But even if we can not assume Gaussian distribution, we nevertheless get:

Proposition 3.26. *(cf. Dacunha Castelle, p. 104, ref in English?) Let X be a STS in L^4 such that $\sup_n E(X_n^4) \leq M^2$ and*

$$\lim_{|n-m|\to\infty} [E[X_nX_{n+k}X_mX_{m+k}] - \gamma^2(k)] = 0.$$

Then $\bar{\gamma}_n(k) \to \gamma(k)$ in L^2.

Proof: : to admit.

(iii) **Comparison** between $\bar{\gamma}$ and γ^*: In the case where $\sup_n E(X_n^4) \leq M^2$ when $n \to \infty$, k being fixed, we get,
Exercise: $\|\bar{\gamma}_n(k) - \gamma_n^*(k)\|_2 \leq \frac{k}{n}M \to 0$.

Routines R: acf, pacf, to give an example.

3.8 ARMA model Identification, estimation of its parameters

Cf. Chapter 5.2 [4].
We assume that the changes in the time series (differentiation, seasonal fitting) have been made so that we have an effective centered STS, and that the obtained series is real, with **a rational spectrum** meaning that there exists p and $q \in \mathbb{N}$, polynomials P degree p and Q degree q, a white noise ε such that the series X is solution to the ARMA equation $A_PX = A_Q\varepsilon$.

The aim is to find p, q, P, Q meaning to identify the model. We have n observations of X and we suppose that the covariance function γ is known, actually estimated according to the method provided in Section 3.7.
R command: arima, monmodele= ; X= ; with model parameters, simulation of processes, plotobs(X) ; mean(X) ; acf(X) which gives correlogram, variogram; pacf(X), etc.

3.8.1 Estimation of P coefficients

Hypothesis: suppose that p, q are known in $\mathbb{N}$ and function γ is known and put $a_0 = 1$.

(p, q) is minimal, meaning there does not exists polynomials P' and Q' with smaller degrees than p, q in the ARMA equation.
We detail the ARMA equation $A_P X = A_Q \varepsilon$:

$$\sum_0^p a_i X_{n-i} = \sum_0^q b_l \varepsilon_{n-l}.$$

We operate the scalar product in L^2 of this equality with X_{n-m} for any $m \geq q+1$, using that X_{n-m} is orthogonal to $(A_Q\varepsilon)_n$, yields for any $m \geq q+1$:

$$\sum_0^p a_i \gamma(m-i) = 0.$$

This is a set of linear equations, the solution of which being the vector a in R^p:

$$\sum_1^p a_i \gamma(m-i) = -\gamma(m), \ \forall m \geq q+1.$$

With $m = q+1, \cdots, q+p$, we get a system of equations named **Yule-Walker equations**; we denote R_{pq} the matrix of this system of p equations and p unknown variables:

$$R_{pq} = \begin{vmatrix} \gamma(q) & \cdots & \gamma(q+1-p) \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ \gamma(q-1+p) & \cdots & \gamma(q) \end{vmatrix}$$

and Γ_{q+1}^{q+p} the vector with coordinates $\gamma(m), \ m = q+1, \cdots, q+p$.

Proposition 3.27. *If X is an ARMA(p,q) process, (p, q) being minimal, the matrix R_{pq} is invertible and the coefficients of the polynomial P are the coordinates of the vector*

$$a = -R_{pq}^{-1} \Gamma_{q+1}^{q+p}.$$

Proof: : to skip, remained for those interested enough.

We assume that $\det R_{pq} = 0$, meaning there exists p coefficients α_i (at least one is non null) such that :

$$\sum_{i=0}^{p-1} \alpha_i \gamma(q+j-i) = 0, \ \forall j = 0, \cdots, p-1.$$

On the other hand, for $j = p$, using Yule-Walker equations, we replace $\gamma(q+p-i)$:

$$\sum_{i=0}^{p-1} \alpha_i \gamma(q+p-i) = -\sum_{i=0}^{p-1} \alpha_i \sum_1^p a_j \gamma(q+p-i-j) = -\sum_{j=1}^p a_j \sum_{i=0}^{p-1} \alpha_i \gamma(q+p-i-j)$$

which is a sum of null terms for $p-j=p-1,\cdots,0$ since $\det R_{pq}=0$. By induction, step by step, we get for $j\geq 0$:

$$\sum_{i=0}^{p-1}\alpha_i\gamma(q+j-i)=0.$$

This exactly reflects the fact that $\forall j\geq 0$:

$$E[\sum_{i=0}^{p-1}\alpha_i X_{n-i}X_{n-j-q}]=0,$$

meaning $\forall n\geq 0, \sum_{i=0}^{p-1}\alpha_i X_{n-i}$ is orthogonal to $H^X_{n-q}=H^\varepsilon_{n-q}$ and we compute its coordinates in $(H^\varepsilon_{n-q})^\perp$:

$$\langle\sum_{i=0}^{p-1}\alpha_i X_{n-i},\varepsilon_{n-q+l}\rangle=\sum_{i=0}^{p-1}\alpha_i\langle X_{n-i},\varepsilon_{n-q+l}\rangle$$

for $l=1,\cdots,q$ and equal to 0 for $l>q$. Moreover using stationarity hypothesis $\langle X_{n-i},\varepsilon_{n-q+l}\rangle$ does not depend on n: since the white noise ε is the innovation white noise X is expressed as a function of ε and this scalar product is stationary.
Denoting γ_l the coordinate of $\sum_{i=0}^{p-1}\alpha_i X_{n-i}$ on ε_{n-q+l}:

$$\sum_{i=0}^{p-1}\alpha_i X_{n-i}=\sum_{l=1}^{q}\gamma_l\varepsilon_{n-q+l},$$

which is an ARMA(p-1,q-1) relation and contradicts the hypothesis that the pair (p,q) is 'minimal'. •

3.8.2 Estimation of Q coefficients

This is a much more difficult problem and we will only give a weak approach! We assume P is known (we estimated it in previous subsection), q and γ are also known. We put

$$Y_n=\sum_{0}^{p}a_k X_{n-k}.$$

We will only put the problem, then its resolution states on numerical analysis. The existence of solutions is proved, but not the uniqueness.
The Y covariance function is computed as a function of the (b_i) using that $Y=A_Q\varepsilon$:

$$\gamma^Y(0)=\textstyle\sum_0^q b_k^2 \quad ;\ \gamma^Y(1)=\sum_1^q b_k b_{k-1}$$
$$\gamma^Y(j)=\textstyle\sum_j^q b_k b_{k-j} \quad ;\ \gamma^Y(q)=b_q b_0$$

We look for a solution b such that the corresponding polynomial Q admits only zeros with modulus ≥ 1.
Exercise: solve this system for $q=1,2$.

For $q=1$, $b_i^2, i=0,1$ are $\frac{1}{2}\left(\gamma(0) \pm \sqrt{\gamma(0)^2 - 4\gamma(1)^2}\right)$ so we need $\gamma(0) \geq 2\gamma(1)$.
For $q=2$..... awful computations !

But the aim is to find the polynomial Q and there is another method, easier but using the complex numbers and what is called "spectral density". Since Y is MA(q) process, its spectral density is known to be

$$f(\lambda) = \frac{1}{2\pi} \sum_{-q}^{+q} \gamma^Y(k) e^{-ik\lambda} = \frac{1}{2\pi} |Q(e^{-i\lambda})|^2$$

where you only have to know that $z = e^{-i\lambda}$ is 2 dimensional, $(\cos(\lambda), -\sin(\lambda))$, and satisfies $1/z = (\cos(\lambda), \sin(\lambda)) = e^{i\lambda}$. So we have to deal with:

$$Q(z)Q(1/z) = \gamma^Y(0) + \sum_{1}^{+q} \gamma^Y(k)(z^k + z^{-k})$$

With the change of variable $Z = z + 1/z$ we compute $z^k + z^{-k}$ as a polynomial of Z, for instance:

$$z^2 + z^{-2} = Z^2 - 2.$$

Thus $Q(z)Q(1/z)$ could be written as a polynomial $U(Z)$ the zero of which, Z_j, are linked to those of Q by the relation $Z_j = z_j + 1/z_j$.

Practically, once found U and its zeros, we deduce those of Q, chosen with modulus ≥ 1. The coefficients b are got from the expansion of $\Pi_j(z - z_j)$.
Remark: CSS= Conditional Square Sum.

Routines R: for instance for ARMA(2,1) needs arima commands:
arima(x, order = c(2,0,1)),
seasonal = list(order = c(2,0,1), period = NA),
xreg = NULL, include.mean = TRUE,
transform.pars = TRUE,
fixed = NULL, init = NULL,
method = c("CSS-ML", "ML", "CSS"), n.cond,
SSinit = c("Gardner1980", "Rossignol2011"),
optim.method = "BFGS",
optim.control = list(), kappa = 1e6)
X.ord=c(2,9,1)
X.arima=arima(X,ord=X.ord)

3.8.3 Characterization of parameters p and q

Definition 3.28. *A rational spectrum ARMA process is said to be with* **minimal type** (p, q) *when in the "canonical Fejer-Riesz relation", the degrees of P and Q are exactly p*

and q.
More concretely: (p, q) is minimal when there does not exist polynomials P' and Q' with smaller degrees than p, q in the ARMA equation.

Consequence: if an ARMA(p',q') process is minimal type (p, q), necessarily $p' \geq p, q' \geq q$.

Theorem 3.29. *A regular STS X is minimal type $(0, q)$ if and only if*

$$\gamma(m) = 0, \ \forall |m| \geq q + 1 \ et \ \gamma(q) \neq 0.$$

Proof: Exercise.
Since $X_n = \sum_{j=0}^{q} b_j \varepsilon_{n-j}$, $\gamma(k) = 0$ as soon as $|k| \geq q + 1$. For $k = q$, $\gamma(q) = a_0 a_q \neq 0$. Conversely, is X is regular, $H_n^X = H_n^\varepsilon$ for any n. The assumption $\gamma(m) = 0, \ \forall |m| \geq q+1$ that X_0 is orthogonal to the space H_{-q-1}^ε. On the other hand, $X_0 \in H_0^\varepsilon$. So $X_0 \in H_0^\varepsilon \cap (H_{-q-1}^\varepsilon)^T$ which is the vector space generated by $\varepsilon_0, ...\varepsilon_{-q}$ so X is MA(q). •

Definition 3.30. *Let (p, q) be a pair of positive numbers. We say that a real series $r_n, \ n \in Z$* **satisfies a** (p, q) **induction** *if there exists coefficients $(\alpha_0, \cdots, \alpha_p)$ with $\alpha_0 = 1, \alpha_p \neq 0$, such that $\sum_0^p \alpha_j r_{m-j} = 0, \ \forall m \geq q + 1$.*

The induction is **minimal** *(p, q) if any pair (p', q') satisfying the property above are such that $p' \geq p, q' \geq q$.*

As we saw that in Subsection 3.8.1, the series $\gamma(n)$ of an ARMA(p,q) satisfies a minimal (p, q) induction. With the γ (or at least their estimates), we can find p and q highlighting the minimal induction. A priori it is not so obvious but this property is equivalent to others properties which are easier to check numerically.

Lemma 3.31. *Let a series $(x_m, \ m \in Z)$ and the matrix $R_{s,t}$ with (i, j) coefficient equal to x_{i-j}, i and j going from 1 to s. If $r_{s,0} \neq 0$, the following are equivalent:*

(i) The series $(x_m, \ m \in Z)$ satisfies a minimal induction (p, q) relation;

(ii) among the determinants $r_{s,t}$, we have $r_{s,t} \neq 0$ while $s \leq p$ or $t \leq q$, and $r_{s,t} = 0$ if $s \geq p + 1$ and $t \geq q + 1$.

(iii) $r_{p+1,q} \neq 0$ and $r_{p,q+1} \neq 0$ and $r_{p+1,j} = 0$ if $j \geq q + 1$.

(iv) $r_{p+1,q} \neq 0$ and $r_{p,q+1} \neq 0$ and $r_{i,q+1} = 0$ if $i \geq p + 1$.

Here $r_{p,q}$ will denote the determinant of the matrix $R_{p,q}$ defined in Section 3.8.1.

Remark 3.32. *In case of ARMA process, $R_{s,0}$ is the variance matrix of the vector $(X_1, \cdots, X_s)$. The lemma hypothesis corresponds to the case where the series X is non singular.*

So this hypothesis is not too strong;
Exercise: if X is non singular, prove that $r_{s,0} \neq 0$. (Meaning: prove that $r_{s,0} = 0$ implies X is singular.)

The lemma proof is tedious, for a complete proof, look at Azencott and Dacunha-Castelle, pp. 137-138.

Proposition 3.33. *Let X be a rational spectrum STS. It is minimal type ARMA (p,q) if and only if the covariance function satisfies a minimal (p,q) induction relation.*
In this case the induction relation is the one which provides the coefficients (a_i) of the polynomial P:

$$\gamma(m) + a_1\gamma(m-1) + \cdots + a_p\gamma(m-p) = 0, \ \forall m \geq q.$$

Definition 3.34. *The* **order s partial auto correlation** *of X, denoted as $\Phi(s)$, is the last coordinate of the vector $-R_{s,0}^{-1}\Gamma_1^s$.*

Previously it was denoted r (Definition 3.2)

$$r(p-n) = \frac{\text{cov } (X_n - X_n^*, X_p - X_p^*)}{\sqrt{\text{Var } (X_n - X_n^*)\ \text{Var } (X_p - X_p^*)}}.$$

Proposition 3.35. *Let a rational spectrum non singular real STS X. It is an AR(p) process if and only if $\Phi(s) = 0, \ \forall s \geq p+1$ and $\Phi(p) \neq 0$.*

Proof: : Necessary condition as an exercise:
Actually X is a regular series and we deal with an innovation white noise process ε.
(i) Remark that by definition $X_n - X_n^*$ is orthogonal to the vector space generated by $\{X_1, ..., X_{n-1}\}$ and X_0^* belongs to this space so $E[(X_n - X_n^*)X_0^*] = 0$. Thus $r(n)$ is proportional to $E[(X_n - X_n^*)X_0]$.
(ii) Since X is AR(p), $X_n = \sum_{j=1}^p a_j X_{n-j} + \varepsilon_n$. Let $n > p$. So $X_n^* = \sum_{j=1}^p a_j X_{n-j}$ since ε_n is orthogonal to the space $\{X_1, ..., X_{p-1}\} \subset H_{n-1}^{\varepsilon}$. Thus $E[(X_n - X_n^*)X_0] = E(\varepsilon_n X_0) = 0$. and $r(n) = 0$ for all $n > p$.
(iii) Finally look at $r(p)$:

$$X_p = \sum_{j=1}^{p} a_j X_{p-j} + \varepsilon_p = \sum_{j=1}^{p-1} a_j X_{p-j} + a_p X_0 + \varepsilon_p \ , \ X_p^* = \sum_{j=1}^{p-1} a_j X_{p-j} + a_p X_0^*, \ X_p - X_p^* = a_p(X_0 - X_0^*) + \varepsilon_p$$

so $E[(X_p - X_p^*)X_0] = a_p E[(X_0 - X_0^*)^2]$ and $r(p) = a_p \neq 0$.

Conversely, to prove the sufficient condition, we use Lemma 3.31. We consider the Cramer system:

$$R_{s,0}\alpha = -\Gamma_1^s.$$

We noticed that, for a non singular series, $r_{s,0} = det\ R_{s,0} \neq 0$. By performing the Cramer resolution, the last coordinate of α is:

$$-\frac{det\ R'_{s,0}}{r_{s,0}}$$

where $R'_{s,0}$ is the matrix $R_{s,0}$ with the last column replaced by Γ_1^s. Using a series of s permutations, we see that $R'_{s,0}$ is actually $R_{s,1}$, and the last coordinate of α is: $(-1)^s \frac{r_{s,1}}{r_{s,0}}$. We then can express the hypothesis

$$\Phi(s) = 0, \ \forall s \geq p+1 \text{ and } \Phi(p) \neq 0$$

as $r_{s,1} = 0\ \forall s \geq p+1$ and $r_{p,1} \neq 0$, meaning the property (iv) in Lemma 3.31 when $q = 0$ which is a characterization of an AR(p). •

References

[1] BOX AND JENKINS : “Times series analysis", Holden-Day, San Francisco, 1976.

[2] R. BROWN : “ Smoothing, Forecasting and Prediction", Prentice Hall, 1962.

[3] R. BROWN : “ Statistical Forecasting for Inventory Control", McGraw Hill, New York, 1959.

[4] C. FRANCQ and J.M. ZAKOIAN, GARCH Models: Structure, Statistical Inference and Financial Applications, Wiley, 2010.

[5] D. COX: “Prediction by exponentially weighted moving average and related methods", Journal of the Royal Statistical Society B,23, p 414-422, 1961.

[6] C. GOURIEROUX et A. MONFORT : “SÃ©ries temporelles et modÃ¨les dynamiques",Economica, Paris,1990.

[7] E.J. HANNAN : “Multiple time series", Wiley, New York, 1970.

[8] P.J. HARRISON : “Exponential Smoothing and short-term sales forecasting" , Management Science, 13 (11), p 821-842, 1967.

[9] C.C. HOLT : “Forecasting Seasonals and Trends by Exponentially Weighted Moving Average", Carnegie Institute of Technology, Pittsburgh, Pennsylvania, 1957.

[10] M. NERLOVE and S. WAGE : “On the optimality of adaptative forecasting" , Management Science, Vol.10,2, p 207-229, 1964.

[11] M. THEIL and S. WAGE : “Some observations on adaptative forecasting" , Management Science, Vol.10,2, p 198-206, 1964.

[12] P. WINTERS : “Forecasting sales by exponentially weighted moving averages" , Management Science, 6, p 324-342, 1960.

[13] T.H. WONNACOTT and R.J. WONNACOTT : "Statistique", Economica, Paris,?.